

Open Source XML Database Toolkit:

*Resources and Techniques
for Improved Development*

Liam Quin

Wiley Computer Publishing



John Wiley & Sons, Inc.

NEW YORK · CHICHESTER · WEINHEIM · BRISBANE · SINGAPORE · TORONTO

Publisher: Robert Ipsen

Editor: Cary Sullivan

Assistant Editor: Christina Berry

Managing Editor: Marnie Wielage

Associate New Media Editor: Brian Snapp

Text Design & Composition: Pronto Design, Inc.

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons, Inc., is aware of a claim, the product names appear in initial capital or ALL CAPITAL LETTERS. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

This book is printed on acid-free paper. ∞

Copyright © 2000 by Liam Quin. All rights reserved.

Published by John Wiley & Sons, Inc.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY, 10158-0012, (212) 850-6011, fax: (212) 850-6008, e-mail: PERMREQ@WILEY.COM.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

Library of Congress Cataloging-in-Publication Data:

Quin, Liam.

Open source XML database toolkit : resources and techniques for improved development / Liam Quin.

p. cm.

ISBN 0-471-37522-5 (pbk. : alk. paper)

1. XML (Document markup language) 2. Web databases. I. Title.

QA76.76.H94 Q56 2000

005.7'2—dc21

00-033022

CIP

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

Acknowledgments	x
Introduction	xi
Part One: Relational Databases and XML	1
Chapter 1: Just Enough XML	3
What Is XML?	3
What XML Is Not	5
XML Reference	6
Additional XML Features	32
Exercises	33
Summary	34
Chapter 2: Client/Server Architecture	35
Client/Server Systems	35
Networks and Protocols	36
Protocols and APIs	37
Asynchronous Networking	45
World Wide Web Architecture	53
Exercises	57
Summary	58
Chapter 3: Just Enough SQL	59
Introduction to Relational Databases	59

The Structured Query Language	63
Normal Forms and Database Design	72
Worked Example: BookWeb	73
Exercises	84
Summary	84
Chapter 4: Generating XML from Relational Data	85
Why Generate XML?	85
Representing Tables	87
Generating XML with DBI	91
Generating XML for a Web Browser: CGI	96
Using PHP	104
Java	112
Exercises	115
Other Approaches	116
Summary	116
Chapter 5: Reading XML into a Program	117
XML Example: Book Catalogue	117
External Parsing and the ESIS	124
Using an Internal Parser	129
SAX: Ælfred Was a Saxon	142
Exercises	142
Summary	143
Chapter 6: XML Database Applications	145
Understanding Requirements	145
XML Architectures	149
Summary	158
Part Two: XML and Databases	159
Chapter 7: What Is a Document?	161

Mixed Content	161
Per-Document Schemata	163
Unrestricted Field Length	163
Arbitrary Field Nesting	164
Field Sequencing	165
Summary	165
Chapter 8: XML Repositories and Databases	167
Sample Scenarios	167
Central Access Scenario	169
Shared Authoring Scenario	170
Revision Control Scenario	171
Information Reuse Scenario	172
Distributed Access and Technology Reuse Scenario	174
Other Issues	175
Summary	175
Chapter 9: Implementation Strategies	177
General Implementation Issues	178
Documents as BLOBs Strategy	180
Paragraphs as BLOBs	183
Elements as Fields Strategy	185
Metadata Only Strategy	187
Elements as Objects Strategy	189
Text Retrieval and Hybrid Approaches	191
Summary	192
Part Three: Nonrelational Databases	193
<hr/>	
Chapter 10: Introduction to Object-Oriented Databases	195
Main Features	195
Summary	203

Chapter 11: XML as Classes and Objects	205
Object Relationships and XML Relationships	206
The Other Way	211
Where Does the Behavior Live?	212
Class Designs	215
Summary	219
Chapter 12: Dynamic Hashing: <i>ndbm</i>	221
What Does <i>ndbm</i> Do?	222
How <i>ndbm</i> Works	224
Using <i>ndbm</i>	225
Performance	236
Using <i>ndbm</i> in Perl	240
Versions of <i>ndbm</i>	242
The <i>ndbm</i> Library and XML	248
Summary	249
Chapter 13: Text Retrieval Technology Overview	251
What Is Text Retrieval?	251
Categorizing Documents	252
Uncategorized Information	253
Queries	254
Multiple Languages	259
Implementation Issues	260
Returning Results to a Program	266
Summary	266
Chapter 14: XQL, XLink, XPath, and XPointer Explained	267
Why So Many Standards?	267
How the Standards Interrelate	268
Related Standards	273
Links and Databases	274
Summary	286

Chapter 15: Hybrid Approaches	287
Files and Databases	287
Databases and Text Retrieval	291
Using <i>ndbm</i> as a Cache Manager	294
Documents as Objects	297
Document Management and Workflow	297
Error Handling	298
Summary	298
Part Four: Links and Metadata	299
<hr/>	
Chapter 16: Just Enough Metadata	301
Metadata Defined	301
History: HTML META and LINK Tags	302
The Dublin Core	304
Other Groups	306
The Resource Description Framework	307
The RDF Schema Specification	311
Who Uses RDF?	311
Other Metadata Standards	312
Summary	313
Chapter 17: Storing Links and Metadata	315
Links as Links and as Metadata	315
Storing Metadata in a Database	316
Groups of Related Documents	321
Extra Link Functionality	321
Link Management and Analysis	324
Document Management	324
Summary	325
Chapter 18: Sketches from the Forge: Sample Applications	327
The AutoLinked Glossary	327

Implementation Overview	332
Other AutoLinker Applications	336
AutoLinker Summary	337
Summary	338
Part Five: Resource Guide	339
<hr/>	
Chapter 19: Open Source Licenses	341
What Is Open Source?	342
The Licenses	345
The GNU General Public License	348
The GNU Lesser General Public License (LGPL)	354
The Artistic License	363
The BSD License	365
The MIT License	366
The Mozilla Public License Version 1.0	366
The Barefoot License	374
Chapter 20: Installing and Configuring Downloaded Software	377
Finding Packages	378
Installing a Binary Package	383
Installing a Source Package	383
Installing a Source Tarball	385
Installing a Perl Module	388
Chapter 21: XML Parsers, Editors, and Utilities	389
Parsers: Tools that Read XML into Memory	389
TeX	394
Browsers	394
Transforming Data	395
Formatting and Printing	396
Editors	396

Chapter 22: Databases, Repositories, and Utilities	401
Relational Databases	401
Object-Oriented Databases	404
Repositories and Document Management Systems	405
Source Repositories	406
Hashing with <i>dbm</i>	409
Hypertext Preprocessor: PHP	409
Perl DBI	409
Information Retrieval Databases	410
Chapter 23: Further Reading	413
Books	413
Magazines and Journals	420
Online Documentation	421
Web Sites	421
Mailing Lists	423
Internet Relay Chat	424
Index	425

ACKNOWLEDGMENTS

This book has been over a year in the making. During that time, the world changed: XML became popular; Perl and Python got XML support; Oracle announced support for Linux; books and magazines on XML, Linux, and Open Source software sprouted like teenagers. Originally, I had thought I would spend most of the book describing tools, but when the tools kept changing, I abandoned that plan, and concentrated on techniques.

Through all this, my editor at Wiley, Christina Berry, remained calm and helpful. I changed jobs; my partner Clyde went back to university; and work on the book dragged on. So thanks are due to Christina, and to Clyde, who insisted that I write this book in the first place. I also need to thank Christopher Cashell for his help with the “Resource Guide”, Jerji (Jerry Herbert) for distracting me when I was close to despair, and, above all, Moonkitty and Cosmos for purring even in the darkest nights.

INTRODUCTION

Welcome to *The Open Source XML Database Toolkit*. This book is not just for people who write code, but also for those who design and specify programs. There are chapters to introduce XML and databases of various types, and chapters to explain how to use them together, all in an open source environment.

Open Source

This book, as is obvious from its title, is about open source software. Most of the tools discussed are open source. Open source software is software that is distributed with its program source code, under a license that allows you to modify and redistribute that source code. In other words, if you don't like the way the program works, *you are free to change it*. People sometimes think that open source software refers to software that is free of charge, but that's not the case. Software in the context of open source is software that you are free to change and to use in any way you wish, as long as you don't try to restrict that freedom from other people.

That said, note that most open source software *is* distributed more or less free of charge, with development costs paid for by support fees, services, or even donations (there may be media and shipping charges, of course).

Also note that this book does mention software that is *not* open source, usually when the software is very significant or when there are no open source competitors. Even if the most appropriate tool is not free, it's better to use it than to be held back by ideology or dogma. Here are some examples of nonfree software mentioned in this book:

Oracle (www.oracle.com). The most widely used high-end commercial relational database.

Solaris (solaris.sun.com). The Sun Solaris operating system running on a Sun SPARC server is probably the best-engineered and most stable server platform today. The source code to Solaris *is* available for a small fee, but the license is restrictive.

SoftQuad XMetal (www.softquad.com). XMetal is a widely used editor for XML documents. It offers a documentlike interface, a structured interface, and a source code view. However, it's only available to run under Microsoft Windows.

Object Design's ObjectStore (www.excelon.com; www.odi.com). ObjectStore is one of the better-known object-oriented databases. Although there is a free version for Java (PSE), source is not available, and there are restrictions on its use.

Most of the software described in this book is free. In a few cases, you may need to pay royalties if you use the software as part of a product or service that you sell, so be sure to check the licenses. Here are a few examples of free software mentioned in this book:

FreeBSD (www.freebsd.org). FreeBSD is one of several open source and free operating systems mentioned in this book. The examples were tested on FreeBSD and Linux.

MySQL (www.mysql.com). MySQL is a freely available relational database, although royalties may apply for some uses. It lacks many of the features of a high-end commercial system such as Oracle's, but it is very widely used and fast, and will take you a long way.

XT (www.jclark.com). XT is an implementation of the XML Style Language Transformation specification, which is a complicated way of saying that it manipulates XML documents, for example to produce HTML or XHTML.

Apache (www.apache.org). Apache is the most widely used Web server on the Internet. In addition to being open source and free, it is also very powerful and robust.

NOTE

Go to www.opensource.org for more information about the open source movement.

XML

The eXtensible Markup Language, XML, is a way of defining simple text-based representations of arbitrarily complex structured information. The term XML is also used to refer to data that's marked up in a format defined using XML.

This is a book about working with XML. You might have XML documents that you need to store in a database, or you might want to use XML as an interchange format.

Chapter 1 in Part One, "Just Enough XML", introduces the main concepts of XML.

Database

This is also book about using databases. You'll find an introduction to the Structured Query Language, SQL, in Chapter 3, "Just Enough SQL."

The book doesn't only address relational databases, though. You'll find descriptions of hashing (Chapter 12: "Dynamic Hashing: *ndbm*"), of object-oriented databases (Chapter 10, "Introduction to Object-Oriented Databases," and Chapter 11, "XML as Classes and Objects"), and of text retrieval databases (Chapter 13, "Text Retrieval Technology Overview").

In all cases, the emphasis is on using XML and databases together in an open source environment.

Toolkit

As in all the best toolkits, there are lots of toys to play with. Most of them are listed for reference in Part Five, "Resource Guide." The toolkit approach means that this book does not go deeply into any single product or tool, but instead focuses on using tools together. Chapter 2, "Client/Server Architecture," introduces network programming, but from then on, the idea of using applications together pervades the book.

The power lies in the way the tools work together. Not only are these open source tools, meaning you can change them to make them work together in the way you want, but they are also widely used and powerful tools, meaning you probably won't have to change them.

Welcome to the open source revolution.

About the Illustrations

The illustrations in printed books generally have to use crisp lines, as if everything was polished and perfect. But don't be deceived by this. A quick sketch on the back of an envelope, or on a whiteboard, can help people to understand the relationships between components where a textual description cannot. Never hesitate to draw pictures, and don't worry if they are not very polished.

Typographic Conventions

I've tried to keep things simple in this regard. In the few places where I've shown a session at an interactive terminal or shell, the prompt is given as the pound sign (`)` if you need to be logged in as `root`, and as the dollar sign (`$`) otherwise. The text you type is in bold. Here's a brief example:

```
$ pwd  
/export/home/liam  
$
```

The dollar sign on the third line shows the prompt after the command (`pwd`, print working directory) completed.

If you are thinking that this looks suspiciously like a Unix (or Linux) shell, you're right. If you're using Microsoft Windows, don't despair—there's a lot you can learn from this book. But if you want to write reliable high-performance database applications, you should develop them on Unix if you can. Imagine going for a whole year of development without a single machine crash and you'll see why.

Source code is shown with function names in **bold** when they are defined. This is just for your convenience, since books don't have a search command. If you type the examples into the computer, or download them, they are plain text, with no formatting. In the same way, comments are shown in *italics*.

What's on the Web Site?

At the companion Web site for this book (www.wiley.com/compbooks/quin) you can find:

- The complete text of the “Resource Guide” in HTML, with links to all of the resources mentioned.
- All of the examples and source code from the book, along with complete or enhanced examples.
- The data for the BookWeb example, along with a simple shell script to create the sample database using MySQL under Linux/Unix.
- The BookWeb site.
- AutoLinker, with the Glossary and the Dictionary examples.
- The sample Web server, written in Perl.

Other XML resources are added from time to time, and the “Resource Guide” is updated occasionally. And note, you may need your copy of the printed book ready before downloading the examples.

Finally, feel free to contact me (liam@holoweb.net; <http://www.holoweb.net/~liam/>); I'm always interested in comments and suggestions for the next edition. You can also find me on the SorceryNet Internet Relay Chat network (irc.sorcery.net; its Web site is at www.sorcery.net).